

A Rewritable, Random-Access DNA-Based Storage System

S. M. Hossein Tabatabaei Yazdi^{1†}, Yongbo Yuan^{2†}, Jian Ma^{3,4}, Huimin Zhao^{2,4},
Olgica Milenkovic^{1*}

Affiliations:

¹Department of Electrical and Computer Engineering, University of Illinois, Urbana, IL 61801

²Department of Chemical and Biomolecular Engineering, University of Illinois, Urbana, IL 61801

³Department of Bioengineering, University of Illinois, Urbana, IL 61801

⁴Institute for Genomic Biology, University of Illinois, Urbana, IL 61801

[†]These authors contributed equally to the work.

^{*}To whom correspondences should be addressed: Olgica Milenkovic, e-mail: milenkov@illinois.edu

Abstract

We describe the first DNA-based storage architecture that enables random access to data blocks and rewriting of information stored at arbitrary locations within the blocks. The newly developed architecture overcomes drawbacks of existing read-only methods that require decoding the whole file in order to read one data fragment. Our system is based on new constrained coding techniques and accompanying DNA editing methods that ensure data reliability, specificity and sensitivity of access, and at the same time provide exceptionally high data storage capacity. As a proof of concept, we encoded parts of the Wikipedia pages of six universities in the USA, and selected and edited parts of the text written in DNA corresponding to three of these schools. The results suggest that DNA is a versatile media suitable for both ultrahigh density archival and rewritable storage applications.

Addressing the emerging demands for massive data repositories, and building upon the rapid development of technologies for DNA synthesis and sequencing, a number of laboratories have recently outlined architectures for archival DNA-based storage [1, 2, 3, 4, 5]. The architecture in [3] achieved a storage density of 700 TB/gram, while the system described in [4] raised the density to 2.2 PB/gram. The success of the latter method may be largely attributed to three *classical coding schemes*: Huffman coding, differential coding, and single parity-check coding [4]. Huffman coding was used for data compression, while differential coding was used for eliminating homopolymers (i.e., repeated consecutive bases) in the DNA strings. Parity-checks were used to add controlled redundancy, which in conjunction with four-fold coverage allows for mitigating assembly errors¹.

Due to dynamic changes in biotechnological systems, none of the three coding schemes represents a suitable solution from the perspective of current DNA sequencer designs: Huffman codes are fixed-to-variable length compressors that can lead to catastrophic error propagation in the presence of sequencing noise; the same is true of differential codes. Homopolymers do not represent a significant source of errors in Illumina sequencing platforms [6], while single parity redundancy or RS codes and differential encoding are inadequate for combating error-inducing sequence patterns such as long substrings with high GC content [6]. As a result, assembly errors are likely, and were observed during the readout process described in [4].

An even more important issue that prohibits the practical wide-spread use of the schemes described in [3, 4] is that accurate partial and random access to data is impossible, as one has to reconstruct the whole text in order to read or retrieve the information encoded even in a few bases. Furthermore, all current designs support read-only storage. The first limitation represents a significant drawback, as one usually needs to accommodate access to specific data sections; the second limitation prevents the use of current DNA storage methods in architectures that call for moderate data editing, for storing frequently updated information and memorizing the history of edits. Moving from a read-only to a rewritable DNA storage system requires a major implementation paradigm shift, as:

¹Another class of DNA error-correcting schemes based on Reed-Solomon (RS) codes was recently reported in [5].

1. Editing in the compressive domain may require rewriting almost the whole information content;
2. Rewriting is complicated by the current data DNA storage format that involves reads of length 100 bps shifted by 25 bps so as to ensure four-fold coverage of the sequence (See Figure 1.1 (a) for an illustration and description of the data format used in [4]). In order to rewrite one base, one needs to selectively access and modify four “consecutive” reads;
3. Addressing methods used in [3, 4] only allow for determining the position of a read in a file, but cannot ensure precise selection of reads of interest, as undesired cross-hybridization between the primers and parts of the information blocks may occur.

To overcome the aforementioned issues, we developed a new, random-access and rewritable DNA-based storage architecture based on DNA sequences endowed with specialized address strings that may be used for selective information access and encoding with inherent error-correction capabilities. The addresses are designed to be *mutually uncorrelated* and to satisfy the *error-control running digital sum constraint* [7, 8]. Given the address sequences, encoding is performed by stringing together properly terminated prefixes of the addresses as dictated by the information sequence. This encoding method represents a special form of *prefix-synchronized coding* [9]. Given that the addresses are chosen to be uncorrelated and at large Hamming distance from each other, it is highly unlikely for one address to be confused with another address or with another section of the encoded blocks. Furthermore, selection of the blocks to be rewritten is made possible by the prefix encoding format, while rewriting is performed via two DNA editing techniques, the gBlock and OE-PCR (overlap-extension polymerase chain reaction) methods [10, 11]. With the latter method, rewriting is done in several steps by using short and cheap primers. The first method is more efficient, but requires synthesizing longer and hence more expensive primers. Both methods were tested on DNA encoded Wikipedia entries of size 17 KB, corresponding to six universities, where information in one, two and three blocks was rewritten in the DNA encoded domain. The rewritten blocks were selected, amplified and Sanger sequenced [12] to verify that selection and rewriting are performed with 100% accuracy.

1 Results

The main feature of our storage architecture that enables highly sensitive random access and accurate rewriting is *addressing*. The rationale behind the proposed approach is that each block in a random access system must be equipped with an address that will allow for unique selection and amplification via DNA sequence primers.

Instead of storing blocks mimicking the structure and length of reads generated during high-throughput sequencing, we synthesized blocks of length 1000 bps tagged at both ends by specially designed address sequences. Adding addresses to short blocks of length 100 bps would incur a large storage overhead, while synthesizing blocks longer than 1000 bps using current technologies is prohibitively costly.

More precisely, each data block of length 1000 bps was flanked at both ends by two unique, yet different, address blocks of length 20 bps. These addresses are used to provide specificity of access (see Figure 1.1 (b) and the Supplementary Information for details). The remaining 960 bases in a block are divided into 12 sub-blocks of length 80 bps, with each block encoding six words of the text. The “word-encoding” process may be seen as a specialized compaction scheme suitable for rewriting, and it operates as follows. First, different words in the text are counted and tabulated in a dictionary. Each word in the dictionary is converted into a binary sequence of length sufficiently long to allow for encoding of the dictionary. For our current implementation and texts of choice, described in the Supplementary Information section, this length was set to 24. Encodings of six consecutive words are subsequently grouped into binary sequences of length 144. The two-bit 11 is appended as a word marker to the left hand side of each binary sequence of length 144, resulting in sequences of length 146 bits. The binary sequences are subsequently translated into DNA blocks of length 80 bps using a new family of DNA prefix-synchronized codes described in the Methods section. Our choice for the number of jointly encoded words is governed by the goal to make rewrites as straightforward as possible and to avoid error propagation due to variable codelengths. Furthermore, as most rewrites include words, rather than individual symbols, the word encoding method represents an efficient means for content update. Details regarding the counting and grouping procedure may be found in the Supplementary Information.

For three selected access queries, the 1000 bps blocks containing the desired information were identified via primers corresponding to their unique addresses, PCR amplified, Sanger sequenced, and subsequently decoded.

Two methods were used for content rewriting. If the region to be rewritten had length exceeding several hundreds, new sequences with unique primers were synthesized as this solution represents a less costly alternative to rewriting. For the case that a relatively short substring of the encoded string had to be modified, the corresponding 1000 bps block hosting the string was identified and the changes were generated via DNA editing.

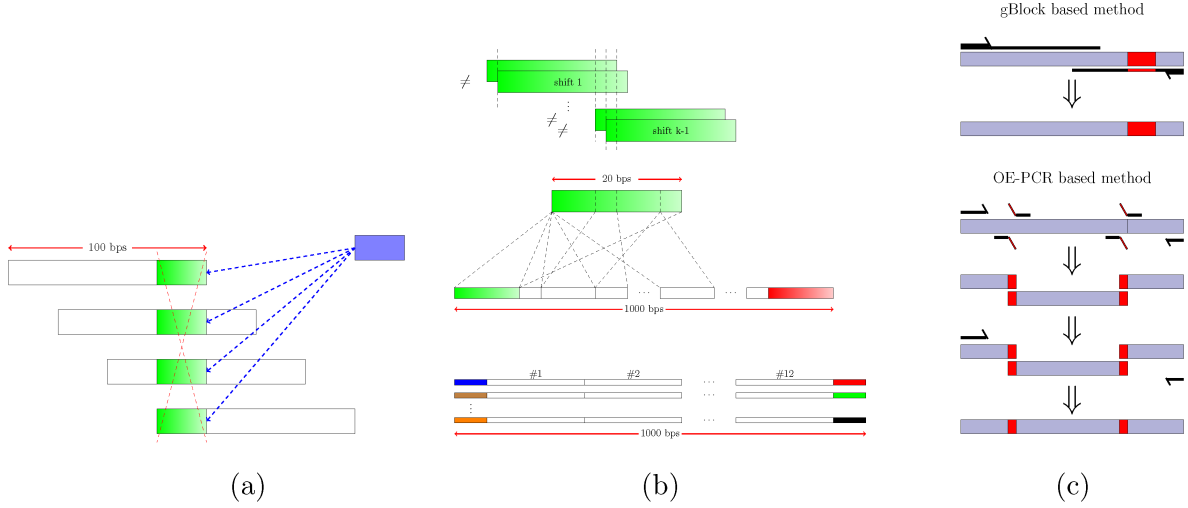


Figure 1.1. (a) The scheme of [4] uses a storage format consisting of DNA strings that cover the encoded compressed text in fragments of length of 100 bps. The fragments overlap in 75 bps, thereby providing 4-fold coverage for all except the flanking end bases. This particular fragmenting procedure prevents efficient file editing: If one were to rewrite the “shaded” block, all four fragments containing this block would need to be selected and rewritten at different positions to record the new “shaded” block. (b) The address sequence construction process using the notions of *autocorrelation and cross-correlation of sequences* [13]. A sequence is uncorrelated with itself if no proper prefix of the sequence is also a suffix of the same sequence. Alternatively, no shift of the sequence overlaps with the sequence itself. Similarly, two different sequences are uncorrelated if no prefix of one sequence matches a suffix of the other. Addresses are chosen to be mutually uncorrelated, and each 1000 bps block is flanked by an address of length 20 on the left and by another address of length 20 on the right (colored ends). (c) Content rewriting via DNA editing: the gBlock method [10] for short rewrites, and the cost efficient OE-PCR (Overlap Extension-PCR) method [11] for sequential rewriting of longer blocks.

Both the random access and rewriting protocols were tested experimentally on two jointly stored text files. One text file, of size 4 KB, contained the history of University of Illinois, Urbana-Champaign (UIUC) based on its Wikipedia entry retrieved on 12/15/2013. The other text file, of size 13 KB, contained the introductory Wikipedia entries of Berkeley, Harvard, MIT, Princeton, and Stanford, retrieved on 04/27/2014.

Encoded information was converted into DNA blocks of length 1000 bps synthesized by IDT (Integrated DNA Technologies), at a cost of \$149 per 1000 bps (see <http://www.idtdna.com/pages/products/genes/gblocks-gene-fragments>). The rewriting experiments encompassed:

1. *PCR selection and amplification of one 1000 bps sequence and simultaneous selection and amplification of three 1000 bps sequences in the pool.* All 32 linear 1000 bps fragments were mixed, and the mixture was used as a template for PCR amplification and selection. The results of amplification were verified by confirming sequence lengths of 1000 bps banks via gel electrophoresis (Figure 1.2 (a)) and by randomly sampling 3-5 sequences from the pools and Sanger sequencing them (Figure 1.2 (b)).

2. *Experimental content rewriting via synthesis of edits located at various positions in the 1000 bps blocks.* For simplicity of notation, we refer to the blocks in the pool on which we performed selection and editing as B1, B2, and B3. Two primers were synthesized for each rewrite in the blocks, for the forward and reverse direction. In addition, two different editing/mutation techniques were used, gBlock and Overlap-Extension (OE) PCR. gBlocks are double-stranded genomic fragments used as primers or for the purpose of genome editing, while OE-PCR is a variant of PCR used for specific DNA sequence editing via point editing/mutations or splicing. To demonstrate the plausibility of a cost efficient method for editing, OE-PCR was implemented with general primers (≤ 60 bps) only. Note that for edits shorter than 40 bps, the mutation sequences were designed as overhangs in primers. Then, the three PCR products were used as templates for the final PCR reaction involving the entire 1000 bps rewrite. Figure 1.1 (c) illustrates the described rewriting process. In addition, a summary of the experiments performed is provided in Table S3.

Sequence identifier - Editing Method	# of sequence samples	Length of edits (bps)	Selection accuracy/error percentage
B1-M-gBlock	5	20	(5/5)/0%
B1-M-PCR	5	20	(5/5)/0%
B2-M-gBlock	5	28	(5/5)/0%
B2-M-PCR	5	28	(5/5)/0%
B3-M-gBlock	5	41 + 29	(5/5)/0%
B3-M-PCR	5	41 + 29	(5/5)/0%

Table 1. Selection, rewriting and sequencing results. Each rewritten 1000 bps sequence was ligated to a linearized pCRTM-Blunt vector using the Zero Blunt PCR Cloning Kit and was transformed into *E. coli*. The *E. coli* strains with correct plasmids were sequenced at ACGT, Inc. Sequencing was performed using two universal primers: M13F_20 (in the reverse direction) and M13R (in the forward direction) to ensure that the entire block of 1000 bps is covered.

	Church et.al. [3]	Goldman et.al. [4]	Our scheme
Density	0.7×10^{15} B/g	2.2×10^{15} B/g	4.9×10^{20} B/g
File size	5.27MB	739KB	File size: 17KB
Cost	Not available	\$12,600	\$4,023
Features	Archival, no random-access	Archival, no random-access	Rewritable, random-access

Table 2. Comparison of storage densities for the DNA *encoded* information expressed in B/g (bytes per gram), file size, synthesis cost, and random access features of three known DNA storage technologies. Note that the density does not reflect the entropy of the information source, as the text files are encoded in ASCII format, which is a redundant representation system.

Given that each nucleotide has weight roughly equal to 650 daltons ($650 \times 1.67 \times 10^{-24}$ grams), and given that $27,000 + 5000 = 32,000$ bps were needed to encode a file of size $13 + 4 = 17$ KB in ASCII format, we estimate a potential storage density of 4.9×10^{20} B/g. This density significantly surpasses the current state-of-the-art storage density of 2.2×10^{15} bytes/g, as we avoid costly multiple coverage, use larger blocklengths and specialized word encoding schemes. A performance comparison of the three currently known DNA-based storage media is given in Table S2. We observe that the cost of sequence synthesis in our storage model is significantly higher than the corresponding cost of the prototype in [4], as blocks of length 1000 bps are still difficult to synthesize. This trend is likely to change dramatically in the near future, as within the last seven months, the cost of synthesizing 1000 bps blocks reduced almost 7-fold. Despite its high cost, our system offers exceptionally large storage density, and for the first time, enables random access and content rewriting features. Furthermore, although we used Sanger sequencing methods for our small scale experiment, for large scale storage projects Next Generation Sequencing (NGS) technologies will enable significant reductions in readout costs.

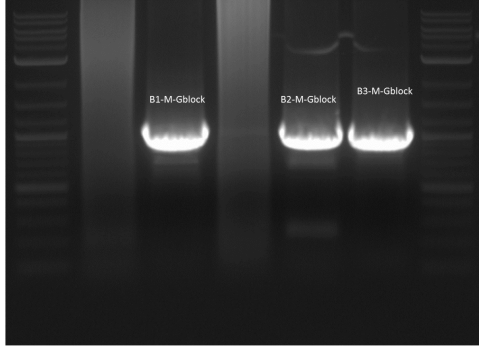
2 Methods

2.1 Address Design and Encoding

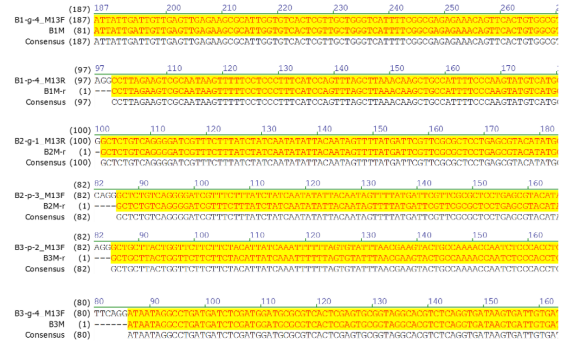
To encode information on DNA media, we employed a two-step procedure. First, we designed address sequences of short length which satisfy a number of constraints that makes them suitable for highly selective random access [13]. *Constrained coding* ensures that DNA patterns prone to sequencing errors are avoided and that DNA blocks are accurately accessed, amplified and selected without perturbing or accidentally selecting other blocks in the DNA pool. The coding constraints apply to address primer design, but also indirectly govern the properties of the fully encoded DNA information blocks. The design procedure used is semi-analytical, in so far that it combines combinatorial methods with computer search techniques.

We required the address sequences to satisfy the following constraints:

- (C1) Constant GC content (close to 50%) of all their prefixes of sufficiently long length. DNA strands with 50% GC content are more stable than DNA strands with lower or higher GC content and have better coverage during sequencing. Since encoding user information is accomplished via prefix-synchronization, it is important to impose



(a)



(b)

Figure 1.2. (a) Gel electrophoresis results for three blocks, indicating that the length of the three selected and amplified sequences is tightly concentrated around 1000 bps. (b) Output of the Sanger sequencer, where all bases shaded in yellow correspond to correct readouts. The sequencing results confirmed that the desired sequences were selected, amplified, and rewritten with 100% accuracy.

the GC content constraint on the addresses as well as their prefixes, as the latter requirement also ensures that all fragments of encoded data blocks have balanced GC content.

- (C2) Large mutual Hamming distance, as it reduces the probability of erroneous address selection. Recall that the Hamming distance between two strings of equal length equals the number of positions at which the corresponding symbols disagree. An appropriate choice for the minimum Hamming distance is equal to half of the address sequence length (10 bps in our current implementation which uses length 20 address primers).
- (C3) Uncorrelatedness of the addresses, which imposes the restriction that prefixes of one address do not appear as suffixes of the same or another address and vice versa. The motivation for this new constraint comes from the fact that addresses are used to provide unique identities for the blocks, and that their substrings should therefore not appear in “similar form” within other addresses. Here, “similarity” is assessed in terms of hybridization affinity. Furthermore, long undesired prefix-suffix matches may lead to read assembly errors in blocks during joint informational retrieval and sequencing.
- (C4) Absence of secondary (folding) structures, as such structures may cause errors in the process of PCR amplification and fragment rewriting.

Addresses satisfying constraints C1-C2 may be constructed via error-correcting codes with small running digital sum [7] adapted for the new storage system. Properties of these codes are discussed in Section 2.2. The novel notion of *mutually uncorrelated sequences* is introduced in 2.3. Constructing addresses that simultaneously satisfy the constraints C1-C4 and determining bounds on the largest number of such sequences is prohibitively complex [14, 15]. To mitigate this problem, we resort to a *semi-constructive* address design approach, in which balanced error-correcting codes are designed independently, and subsequently expurgated so as to identify a large set of mutually uncorrelated sequences. The resulting sequences are subsequently tested for secondary structure using *mfold* and *Vienna* [16]. We conjecture that the number of sequences satisfying C1-C4 grows exponentially with their length: proofs towards establishing this claim include results on the exponential size of codes under each constraint individually.

Given two uncorrelated sequences as flanking addresses of one block, one of the sequences is selected to encode user information via a new implementation of *prefix-synchronized encoding* [17, 16], described in 2.4. The asymptotic rate of an optimal single sequence prefix-free codes is one. Hence, there is no asymptotic coding loss for avoiding prefixes of one sequence; we only observe a minor coding loss for each finite-length block. For multiple sequences of arbitrary structure, the problem of determining the optimal code rate is significantly more complicated and the rates have to be evaluated numerically, by solving systems of linear equations [17] as described in 2.4 and the Supplementary Information. This system of equations leads to a particularly simple form for the generating function of mutually uncorrelated sequences, as explained in the Supplementary Information.

2.2 Balanced Codes and Running Digital Sums

An important criteria for selecting block addresses is to ensure that the corresponding DNA primer sequences have prefixes with a GC content approximately equal to 50%, and that the sequences are at large pairwise Hamming distance. Due to their applications in optical storage, codes that address related issues have been studied in a different form under the name of *bounded running digital sum* (BRDS) codes [7, 8]. A detailed overview of this coding technique may be found in [7].

Consider a sequence $a = a_0, a_1, a_2, \dots, a_l, \dots, a_n$ over the alphabet $\{-1, 1\}$. We refer to $S_l(a) = \sum_{i=0}^{l-1} a_i$ as the running digital sum (RDS) of the sequence a up to length l , $l \geq 0$. Let $D_a = \max\{|S_l(a)| : l \geq 0\}$ denote the largest value of the running digital sum of the sequence a . For some predetermined value $D > 0$, a set of sequences $\{a(i)\}_{i=1}^M$ is termed a BRDS code with parameter D if $D_{a(i)} \leq D$ for all $i = 1, \dots, M$. Note that one can define non-binary BRDS codes in an equivalent manner, with the alphabet usually assumed to be symmetric, $\{-q, -q+1, \dots, -1, 1, \dots, q-1, q\}$, and where $q \geq 1$. A set of DNA sequences over $\{\mathbf{A}, \mathbf{T}, \mathbf{G}, \mathbf{C}\}$ may be constructed in a straightforward manner by mapping each $+1$ symbol into one of the bases $\{\mathbf{A}, \mathbf{T}\}$, and -1 into one of the bases $\{\mathbf{G}, \mathbf{C}\}$, or vice versa. Alternatively, one can use BRDS over an alphabet of size four directly.

To address the constraints C1-C2, one needs to construct a large set of BRDS codewords at sufficiently large Hamming distance from each other. Via the mapping described above, these codewords may be subsequently translated to DNA sequences with a GC content approximately equal to 50% for all sequence prefixes, and at the same Hamming distance as the original sequences.

Let $(n, C, d; D)$ be the parameters of a BRDS error-correcting code, where C denotes the number of codewords of length n , d denotes the minimum distance of the code, while $\frac{\log C}{n}$ equals the code rate. For $D = 1$ and $d = 2$, the best known BRDS-code has parameters $(n, 2^{\frac{n}{2}}, 2; 1)$, while for $D = 2$ and $d = 1$, codes with parameters $(n, 3^{\frac{n}{2}}, 1; 2)$ exist. For $D = 2$ and $d = 2$, the best known BRDS code has parameters $(n, 2 \cdot 3^{(\frac{n}{2})-1}, 2; 2)$ [8]. Note that each of these codes has an exponentially large number of codewords, among which a (sufficiently) large number of sequences satisfy the required correlation property C3, discussed next, and the folding property C4. Codewords satisfying constraints C3-C4 were found by expurgating the BRDS codes via computer search.

2.3 Sequence Correlation

We describe next the notion of autocorrelation of a sequence and introduce the related notion of mutual correlation of sequences.

It was shown in [17] that the autocorrelation function is the crucial mathematical concept for studying sequences avoiding forbidden strings and substrings. In the storage context, forbidden strings correspond to the addresses of the blocks in the pool. In order to accommodate the need for selective retrieval of a DNA block without accidentally selecting any undesirable blocks, we find it necessary to also introduce the notion of mutually uncorrelated sequences.

Let X and Y be two words, possibly of different lengths, over some alphabet of size $q > 1$. The correlation of X and Y , denoted by $X \circ Y$, is a binary string of the same length as X . The i -th bit (from the left) of $X \circ Y$ is determined by placing Y under X so that the leftmost character of Y is under the i -th character (from the left) of X , and checking whether the characters in the overlapping segments of X and Y are identical. If they are identical, the i -th bit of $X \circ Y$ is set to 1, otherwise, it is set to 0. For example, for $X = \text{CATCATC}$ and $Y = \text{ATCATCGG}$, $X \circ Y = 0100100$, as depicted below.

Note that in general, $X \circ Y \neq Y \circ X$, and that the two correlation vectors may be of different lengths. In the example above, we have $Y \circ X = 00000000$. The autocorrelation of a word X equals $X \circ X$.

In the example below, $X \circ X = 1001001$.

$$\begin{array}{ccccccccccccccccc} X = & C & A & T & C & A & T & C & & & & & & & & & \\ Y = & A & T & C & A & T & C & G & G & & & & & & & & 0 \\ & & A & T & C & A & T & C & G & G & & & & & & & 1 \\ & & & A & T & C & A & T & C & G & G & & & & & & 0 \\ & & & & A & T & C & A & T & C & G & G & & & & & 0 \\ & & & & & A & T & C & A & T & C & G & G & & & & 1 \\ & & & & & & A & T & C & A & T & C & G & G & & & 0 \\ & & & & & & & A & T & C & A & T & C & G & G & & 0 \end{array}$$

Definition 1. A sequence X is *self-uncorrelated* if $X \circ X = 10 \dots 0$. A set of sequences $\{X_1, X_2, \dots, X_m\}$ is termed *mutually uncorrelated* if each sequence is self-uncorrelated and if all pairs of distinct sequences satisfy $X_i \circ X_j = 0 \dots 0$ and $X_j \circ X_i = 0 \dots 0$.

Intuitively, correlation captures the extent to which prefixes of sequences overlap with suffixes of the same or other sequences. Furthermore, the notion of mutual uncorrelatedness may be relaxed by requiring that only sufficiently long prefixes do not match sufficiently long suffixes of other sequences. Sequences with this property, and at sufficiently large Hamming distance, eliminate undesired address cross-hybridization during selection and cross-sequence assembly errors.

We proved the following bound on the size of the largest mutually uncorrelated set of sequences of length n over an alphabet of size $q = 4$. The bounds show that there exist exponentially many mutually uncorrelated sequences for any choice of n , and the lower bound is constructive. Furthermore, the construction used in the bound “preserves” the Hamming distance (see the Supplementary Information).

Theorem 2. Suppose that $\{X_1, \dots, X_m\}$ is a set of m pairwise mutually uncorrelated sequences of length n . Let $u(n)$ denote the largest possible value of m for a given n . Then

$$4 \cdot 3^{\frac{n}{4}} \leq u(n) \leq 9 \cdot 4^{n-2}.$$

As an illustration, for $n = 20$, the *lower bound* equals 972. The proof of the theorem is give in the Supplementary Information.

It remains an open problem to determine the largest number of address sequences that jointly satisfy the constraints C1-C4. We conjecture that the number of such sequences is exponential in n , as the numbers of words that satisfy C1-C2, C3 and C4 [15] are exponential. Exponentially large families of address sequences are important indicators of the scalability of the system and they also influence the rate of information encoding in DNA.

Using a casting of the address sequence design problem in terms of a simple and efficient greedy search procedure, we were able to identify 1149 sequences for length $n = 20$ that satisfy constraints C1-C4, out of which 32 pairs were used for block addressing. Another means to generate large sets of sequences satisfying the constraints is via approximate solvers for the *largest independent set problem* [18]. Examples of sequences constructed in the aforementioned manner and used in our experiments are listed in the Supplementary Information.

2.4 Prefix-Synchronized DNA Codes

In the previous sections, we described how to construct address sequences that can serve as unique identifiers of the blocks they are associated with. We also pointed out that once such address sequences are identified, user information has to be encoded in order to *avoid* the appearance of any of the addresses, sufficiently long substrings of the addresses, or substrings similar to the addresses in the resulting DNA codeword blocks. For this purpose, we developed new prefix-synchronized encoding schemes based on [14].

To address the problem at hand, we start by introducing comma free and prefix-synchronized codes which allow for constructing codewords that avoid address patterns. A block code \mathcal{C} comprising a set of codewords of length N over an alphabet of size q is called *comma free* if and only if for any pair of not necessarily distinct codewords $a_1 a_2 \dots a_N$ and $b_1 b_2 \dots b_N$ in \mathcal{C} , the N concatenations $a_2 a_3 \dots a_N b_1, a_3 a_4 \dots b_1 b_2, \dots, a_N a_1 \dots b_{N-2} b_{N-1}$ are not in \mathcal{C} [17]. Comma free codes enable efficient synchronization protocols, as one is able to determine the starting positions of codewords without ambiguity. A major drawback of comma free codes is the need to implement an exhaustive search procedure over sequence sets to decide whether or not a given string of length n should be used as a codeword or not. This difficulty can be overcome by using a special family of comma free codes, introduced by Gilbert [9] under the name *prefix-synchronized codes*. Prefix-synchronized codes have the property that every codeword starts with a prefix $P = p_1 p_2 \dots p_n$, which is followed by a constrained sequence $c_1 c_2 \dots c_s$. Moreover, for any codeword $p_1 p_2 \dots p_n c_1 c_2 \dots c_s$ of length $n + s$, the prefix P *does not appear* as a substring of $p_2 \dots p_n c_1 c_2 \dots c_s p_1 p_2 \dots p_{n-1}$. More precisely, the constrained sequences of prefix-synchronized codes avoid the pattern P which is used as the address.

Due to the choice of mutually uncorrelated addresses at *large Hamming distance*, we encode each information block by *avoiding only one of the address sequences*, used for that particular block.

To explain how to perform encoding, assume that $P = p_1 p_2 \dots p_n \in \{\mathbf{A}, \mathbf{T}, \mathbf{G}, \mathbf{C}\}^n$ is a self-uncorrelated sequence. This guarantees that $p_1 \neq p_n$. Without loss of generality, let $p_1 = \mathbf{A}$ and $p_n = \mathbf{G}$, and define

$$\begin{aligned} \bar{P}_i &= \{\mathbf{A}, \mathbf{C}, \mathbf{T}\} \setminus \{p_i\} \\ P^i &= p_1 \dots p_i, \end{aligned}$$

for all $1 \leq i \leq n$. In addition, assume that the elements of \bar{P}_i are arranged in increasing order, say using the lexicographical ordering $A < C < T$. We subsequently use $\bar{p}_{i,j}$ to denote the j -th smallest element in \bar{P}_i , for $1 \leq j \leq |\bar{P}_i|$. For example, if $\bar{P}_i = \{C, T\}$, then $\bar{p}_{i,1} = C$ and $\bar{p}_{i,2} = T$.

Next, we define a sequence of integers $G_{n,1}, G_{n,2}, \dots$ that satisfies the following recursive formula

$$G_{n,\ell} = \begin{cases} 3^\ell, & 1 \leq \ell < n, \\ \sum_{i=1}^{n-1} |\bar{P}_i| G_{n,\ell-i}, & \ell \geq n. \end{cases}$$

For an integer $\ell \geq 0$ and $y < 3^\ell$, let $\theta_\ell(y) = \{A, T, C\}^\ell$ be a length- ℓ ternary representation of y . Conversely, for each $W \in \{A, T, C\}^\ell$, let $\theta^{-1}(W)$ be the integer y such that $\theta_\ell(y) = W$. Every integer $0 \leq x < G_{n,\ell}$ can be mapped into a sequence of $n + \ell$ symbols $\{A, T, C, G\}$ via an encoding algorithm that consists of two parts: **EncodePSC**(P, ℓ, x) and **CodePSC**(P, ℓ, x). Algorithm **EncodePSC**(P, ℓ, x) calls **CodePSC**(P, ℓ, x) and returns the concatenation of P and **CodePSC**(P, ℓ, x).

The steps of the encoding procedure are listed in Algorithm 1, where $C_\ell^P = \{\text{EncodePSC}(P, \ell, x) \mid 0 \leq x < G_{n,\ell}\}$, and where n denotes the length of the sequence P . The decoding steps are described in the same chart.

Algorithm 1 Prefix-synchronized encoding and decoding

$X = \text{EncodePSC}(P, \ell, x)$

return $P\text{CodePSC}(P, \ell, x)$;

$X = \text{CodePSC}(P, \ell, x)$

begin

1 $n = \text{length}(P)$;

2 if $(\ell \geq n)$

3 $t := 1$;

4 $y := x$;

5 while $(y \geq |\bar{P}_t| G_{n,\ell-t})$

6 $y := y - |\bar{P}_t| G_{n,\ell-t}$;

7 $t++$;

8 end;

9 $a := \lfloor \frac{y}{G_{n,\ell-t}} \rfloor$;

10 $b := \text{mod}(y, G_{n,\ell-t})$;

11 return $P^{t-1} \bar{p}_{t,a+1} \text{CodePSC}(P, \ell-t, b)$;

12 else

13 return $\theta_\ell(y)$;

14 end;

end;

$x = \text{DecodePSC}(P, X)$

begin

1 $n = \text{length}(P)$;

2 $\ell = \text{length}(X)$;

3 $X = X_1 X_2 \dots X_\ell$;

4 if $(\ell < n)$

5 return $\theta^{-1}(X)$;

6 else

7 find (s, t) such that $P^{t-1} \bar{p}_{t,s} = X_1 \dots X_t$;

8 return $(\sum_{i=1}^{t-1} |\bar{P}_i| G_{n,\ell-i}) + (s-1) G_{n,\ell-t} + \text{DecodePSC}(P, X_{t+1} \dots X_\ell)$;

9 end;

end;

The following theorems are proved in the Supplementary Information.

Theorem 3. C_ℓ^P is a prefix-synchronized codeword.

Theorem 4. The algorithm **EncodePSC**(P, ℓ, x) outputs a uniquely decodable string, for any $0 \leq x < G_{n,\ell}$.

A simple example describing the encoding and decoding procedure for the short address string $P = \text{AGCTG}$, which can easily be verified to be self-uncorrelated, is provided in the Supplementary Information.

The previously described **EncodePSC**(P, ℓ, x) algorithm imposes no limitations on the length of a prefix used for encoding. This feature may lead to unwanted cross hybridization between address primers used for selection and the prefixes of addresses encoding the information. One approach to mitigate this problem is to “perturb” long prefixes in the encoded information in a controlled manner. For small-scale random access/rewriting experiments, the recommended approach is to first select all prefixes of length greater than some predefined threshold. Afterwards, the first and last quarter of the bases of these long prefixes are used unchanged while the central portion of the prefix string is cyclically shifted by half of its length. For example, for the address primer **ACTAACTGTGCGACTGATGC**, the suffix **ACTAACTGTGCGACTG** produced by **EncodePSC**(P, ℓ, x) maps to **ACTAATGCCTGGACTG**. The process of shifting applied to this string is illustrated below:

ACTAA CTGTGC GACTG
 cyclically shift by 3
 ↓
 ACTAA TGCCTG GACTG

For an arbitrary choice of the addresses, this scheme may not allow for unique decoding $\text{EncodePSC}(P, \ell, x)$. However, there exist simple conditions that can be checked to eliminate primers that do not allow this transform to be “unique”. Given the address primers created for our random access/rewriting experiments, we were able to uniquely map each modified prefix to its original prefix and therefore uniquely decode the readouts.

As a final remark, we would like to point out that prefix-synchronized coding also supports error-detection and limited error-correction. Error-correction is achieved by checking if each substring of the sequence represents a prefix or “shifted” prefix of the given address sequence and making proper changes when needed.

3 Discussion

We described a new DNA based storage architecture that enables accurate random access and cost-efficient rewriting. The key component of our implementation is a new collection of coding schemes and the adaptation of random-access enabling codes from classical storage systems. In particular, we encoded information within blocks with unique addresses that are prohibited to appear anywhere else in the encoded information, thereby removing any undesirable cross-hybridization problems during the process of selection and amplification. We also performed four access and rewriting experiments without readout errors, as confirmed by post-selection and rewriting Sanger sequencing. The current drawback of our scheme is high cost, as synthesizing long DNA blocks is expensive. Cost considerations also limited the scope of our experiments and the size of the prototype, as we aimed to stay within a budget comparable to that used for other existing architectures. Nevertheless, the benefits of random access and other unique features of the proposed system compensate for this high cost, which we predict will decrease rapidly in the very near future.

4 acknowledgments

This work was partially supported by the Strategic Research Initiative of University of Illinois, Urbana-Champaign, and the NSF STC on Science of Information, Purdue University. A provisional patent for rewritable, random-access DNA-based storage was filed with the University of Illinois in November 2014.

References

- [1] C. Bancroft, T. Bowler, B. Bloom, and C. T. Clelland, “Long-term storage of information in dna.” *Science (New York, NY)*, vol. 293, no. 5536, pp. 1763–1765, 2001.
- [2] J. Davis, “Microvenus,” *Art Journal*, vol. 55, no. 1, pp. 70–74, 1996.
- [3] G. M. Church, Y. Gao, and S. Kosuri, “Next-generation digital information storage in dna,” *Science*, vol. 337, no. 6102, pp. 1628–1628, 2012.
- [4] N. Goldman, P. Bertone, S. Chen, C. Dessimoz, E. M. LeProust, B. Sipos, and E. Birney, “Towards practical, high-capacity, low-maintenance information storage in synthesized dna,” *Nature*, 2013.
- [5] R. N. Grass, R. Heckel, M. Puddu, D. Paunescu, and W. J. Stark, “Robust chemical preservation of digital information on dna in silica with error-correcting codes,” *Angewandte Chemie International Edition*, vol. 54, no. 8, pp. 2552–2555, 2015.
- [6] M. G. Ross, C. Russ, M. Costello, A. Hollinger, N. J. Lennon, R. Hegarty, C. Nusbaum, and D. B. Jaffe, “Characterizing and measuring bias in sequence data,” *Genome Biol.*, vol. 14, no. 5, p. R51, 2013.
- [7] G. D. Cohen and S. Litsyn, “Dc-constrained error-correcting codes with small running digital sum,” *Information Theory, IEEE Transactions on*, vol. 37, no. 3, pp. 949–955, 1991.
- [8] M. Blaum, S. Litsyn, V. Buskens, and H. C. van Tilborg, “Error-correcting codes with bounded running digital sum,” *IEEE transactions on information theory*, vol. 39, no. 1, pp. 216–227, 1993.
- [9] E. Gilbert, “Synchronization of binary messages,” *Information Theory, IRE Transactions on*, vol. 6, no. 4, pp. 470–477, 1960.
- [10] H. Packer, “gblocks® gene fragments, related decoded articles,” 2014.

- [11] A. V. Bryksin and I. Matsumura, “Overlap extension pcr cloning: a simple and reliable way to create recombinant plasmids,” *Biotechniques*, vol. 48, no. 6, p. 463, 2010.
- [12] S. C. Schuster, “Next-generation sequencing transforms today’s biology,” *Nature methods*, vol. 5, no. 1, pp. 16–18, 2008.
- [13] K. A. S. Immink, *Codes for mass data storage systems*. Shannon Foundation Publisher, 2004.
- [14] H. Morita, A. J. van Wijngaarden, and A. Han Vinck, “On the construction of maximal prefix-synchronized codes,” *Information Theory, IEEE Transactions on*, vol. 42, no. 6, pp. 2158–2166, 1996.
- [15] O. Milenkovic and N. Kashyap, “On the design of codes for dna computing,” in *Coding and Cryptography*. Springer, 2006, pp. 100–119.
- [16] J.-M. Rouillard, M. Zuker, and E. Gulari, “Oligoarray 2.0: design of oligonucleotide probes for dna microarrays using a thermodynamic approach,” *Nucleic acids research*, vol. 31, no. 12, pp. 3057–3062, 2003.
- [17] L. J. Guibas and A. M. Odlyzko, “Maximal prefix-synchronized codes,” *SIAM Journal on Applied Mathematics*, vol. 35, no. 2, pp. 401–418, 1978.
- [18] P. Berman and M. Fürer, “Approximating maximum independent set in bounded degree graphs.” in *SODA*, vol. 94, 1994, pp. 365–371.
- [19] R. G. Gallager, “Low-density parity-check codes,” *Information Theory, IRE Transactions on*, vol. 8, no. 1, pp. 21–28, 1962.

Supplementary Information

List of sections

1. Encoding Wikipedia Entries – A Working Example (Section 1).
2. Proofs of Theorems (Section 2).
3. Address Sequences (Section 3).
4. Example of Encoding and Decoding Procedure (Section 4).
5. Experimental Synthesis, Access and Rewrite of DNA Storage Sequences (Section 5).
6. Hybrid DNA-Based and Classical Storage (Section 6).

1 Encoding Wikipedia entries: A Working Example

In this section we describe the data format used for encoding two files of size 17 KB containing the introductory sections of Wikipedia pages of six universities: Berkeley, Harvard, MIT, Princeton, Stanford, and University of Illinois Urbana-Champaign. There were 1,933 words in the text, out of which 842 were distinct. Note that in our context, words are elements of the text separated by a space. For example, “university” and “university.” are counted as two different words, while “Urbana-Champaign” is counted as a single word. These 1,933 words were mapped to $\lceil \frac{1933}{72} \rceil = 27$ DNA blocks of length 1000 bps, as we grouped six words into fragments, and combined 12 fragments for prefix-synchronized encoding. Table S1 provides the word counts in the files and encoding lengths (in bits) of the of the outlined procedure.

Assume that instead of using a prefix-synchronized code, we used classical ASCII encoding without compression to encode the same Wikipedia pages. The total number of characters in the text equals 12,874, and each character is mapped to a binary string of length 7. Hence, one would need $12874 \times 7 = 90118$ bits to represent the data, which is equivalent to $\lceil \frac{90118}{2 \times 960} \rceil = 47$ DNA blocks of length 1000 bps if we set aside two unique address flags for the blocks. As one can see, prefix-synchronized codes offer an almost 1.7-fold improvement in description length compared to ASCII encoding. This comes at the cost of storing a larger dictionary, as one encodes words rather than symbols of the alphabet. For the working example, one would require roughly 70-times larger dictionaries, as there are 1933 words with an average of 5.1 symbols per word. This increased in the dictionary is not a significant problem, as only one copy of the dictionary is ever needed.

	# symbols	# distinct symbols	# bits/distinct symbol	# bits
Characters	12874	51	6	77244
Words	1933	842	12	23196

Table S1. Comparison between character and word based encoding. Note the the number of bits per distinct symbol for the word encoding case is computed as the ceiling of the logarithm of the number of distinct symbols plus one, where the extra bit is used to prevent very small integers from being used in prefix-synchronized coding. Such integers may produce long runs of the first symbol in the address, which should be avoided. Furthermore, to ensure fixed length encoding, and hence avoid catastrophic error propagation, we doubled the number of bits used for encoding to 24.

2 Proofs of Theorems

Proof of Theorem 2. The proof consists of two parts. First, we prove the upper bound on $u(n)$ in Lemma 1, and then proceed to prove a lower bound in Lemma 2. Recall that $u(n)$ denotes the largest possible size for a set of mutually uncorrelated words of length n .

Lemma 1. *Let $u(n)$ the largest set of distinct mutually uncorrelated sequences of length n . Then*

$$u(n) \leq 9 \cdot 4^{n-2}.$$

Proof: To prove the lemma, let us introduce some terminology. Let $d_H(\cdot, \cdot)$ stand for the Hamming distance between two words, and define the Hamming ball of radius d around a point W in $\{\mathbf{A}, \mathbf{T}, \mathbf{G}, \mathbf{C}\}^n$ as

$$B(W, d) = \{W' \in \{\mathbf{A}, \mathbf{T}, \mathbf{G}, \mathbf{C}\}^n : d_H(W, W') \leq d\}.$$

Furthermore, let

$$C(W, d) = \{W' \in \{\mathbf{A}, \mathbf{T}, \mathbf{G}, \mathbf{C}\}^n : W' \in B(W, d), W', W \text{ are correlated}\}$$

denote the set of sequences correlated with W that are also at most at Hamming distance d from W .

We claim that for $n \geq d + 2 \geq 4$, one has

$$|C(W, d)| \geq 2 \sum_{i=0}^{d-1} \binom{n-1}{i} 3^i - \sum_{i=0}^{d-2} \binom{n-2}{i} 3^i. \quad (2.1)$$

To prove the result, assume without loss of generality that W starts with the symbol \mathbf{A} , i.e., $W = \mathbf{A}W_2 \dots W_n$. Next, consider two scenarios regarding the structure of $W = \mathbf{A}W_2 \dots W_n$:

- $W_n \neq \mathbf{A}$: In this case, any word W' in $B(W, d)$ that starts with W_n or ends with \mathbf{A} is an element of $C(W, d)$.

Let $S = \{W' : W' \in B(W, d), W' \text{ starts with } W_n\}$ and $E = \{W' : W' \in B(W, d), W' \text{ ends with } \mathbf{A}\}$.

Clearly, $|S| = |E| = \sum_{i=0}^{d-1} \binom{n-1}{i} 3^i$ and $|S \cap E| = \sum_{i=0}^{d-2} \binom{n-2}{i} 3^i$. Therefore, $|C(W, d)| \geq |S \cup E| = 2 \sum_{i=0}^{d-1} \binom{n-1}{i} 3^i - \sum_{i=0}^{d-2} \binom{n-2}{i} 3^i$.

- $W_n = \mathbf{A}$: In this case, any word W' in $B(W, d)$ which starts or ends with \mathbf{A} is also an element of $C(W, d)$. Using an argument similar to the one described for the previous scenario, one can show that $|C(W, d)| \geq 2 \sum_{i=0}^d \binom{n-1}{i} 3^i - \sum_{i=0}^d \binom{n-2}{i} 3^i$.

Moreover, it is straightforward to see that

$$2 \sum_{i=0}^d \binom{n-1}{i} 3^i - \sum_{i=0}^d \binom{n-2}{i} 3^i > 2 \sum_{i=0}^{d-1} \binom{n-1}{i} 3^i - \sum_{i=0}^{d-2} \binom{n-2}{i} 3^i.$$

For any mutually uncorrelated set $\{X_1, \dots, X_m\}$ of size m , we have $X_i \notin C(X_1, n)$, for $2 \leq i \leq m$. This implies that

$$\{X_1, \dots, X_m\} \subseteq \{\mathbf{A}, \mathbf{T}, \mathbf{C}, \mathbf{G}\}^n \setminus C(X_1, n).$$

At the same time, the previous claim suggests that

$$\begin{aligned} |C(X_1, n)| &\geq 2 \sum_{i=0}^{n-1} \binom{n-1}{i} 3^i - \sum_{i=0}^{n-2} \binom{n-2}{i} 3^i \\ &= 2 \cdot 4^{n-1} - 4^{n-2}. \end{aligned}$$

Therefore, $m \leq 4^n - (2 \cdot 4^{n-1} - 4^{n-2}) = 9 \cdot 4^{n-2}$, which completes the proof.

Lemma 2. *Let $u(n)$ the largest set of distinct mutually uncorrelated sequences of length n . Then*

$$u(n) \geq 4 \cdot 3^{\frac{n}{4}}.$$

Proof: For simplicity, assume that m is even. Given a mutually uncorrelated set $\{X_1, \dots, X_m\}$, with words of length n and over the alphabet $\{\mathbf{A}, \mathbf{T}, \mathbf{G}, \mathbf{C}\}$, partition $\{X_1, \dots, X_m\}$ into two arbitrary sets A and B of equal size, say $A = \{X_1, \dots, X_{\frac{m}{2}}\}$ and $B = \{X_{\frac{m}{2}+1}, \dots, X_m\}$. We argue that $C = \{XY \mid X \in A, Y \in B\}$ is a mutually uncorrelated set with words of length $2n$.

- First, we show that the elements in C are self-uncorrelated: For an arbitrary element $Z \in C$, we have $Z = XY$. Since the two sequences $\{X, Y\}$ are mutually uncorrelated, one can easily verify that $Z_1^i \neq Z_{2n-i+1}^{2n-i}$, for $i \in \{1, \dots, 2n-1\} \setminus \{n\}$. Moreover, since $X \neq Y$, it holds that $Z_1^n \neq Z_{n+1}^n$. This establishes the claim.
- Next, we argue that any two distinct elements in C are uncorrelated: For any two distinct elements $Z = XY$ and $Z' = X'Y'$ in C , one can show that $Z_1^i \neq (Z')_{2n-i+1}^{2n-i}$, for $i \in \{1, \dots, 2n-1\} \setminus \{n\}$. In addition, $X \neq Y'$ implies that $Z_1^n \neq (Z')_{n+1}^n$. This completes the proof.

As a result, given a mutually uncorrelated set $\{X_1, \dots, X_m\}$, where $X_i \in \{\mathbf{A}, \mathbf{T}, \mathbf{C}, \mathbf{G}\}^n$, one can construct another mutually uncorrelated set $\{Z_1, \dots, Z_{\frac{m}{2}}\}$, where $Z_i \in \{\mathbf{A}, \mathbf{T}, \mathbf{C}, \mathbf{G}\}^{2n}$. Therefore, $u(2n) \geq \frac{u^2(n)}{4}$. Observing that for $n = 4$ it is possible to construct the following set of 12 mutually uncorrelated sequences

$$\begin{aligned} &\{\text{ATGC, ATAC, GTAC, GTGC} \\ &\text{ATTC, GTTC, AGGC, AAAC} \\ &\text{GAAC, GGGC, ATTT, GTTT}\} \end{aligned}$$

establishes the base of a recursive procedure which gives $u(n) > 4 \cdot (1.31)^n$. Note that this bound is constructive, and the concatenation procedure preserves normalized minimum Hamming distances.

We now turn our attention to prefix-synchronized coding, and describe a number of results relevant for our subsequent discussion.

Theorem 5 ([17]). *Given a positive integer N , chose the unique integer $n = n(N)$ so that $\beta = N2^{-n}$ satisfies*

$$\log 2 \leq \beta < 2 \log 2.$$

Then, the maximal prefix-synchronized code of length N has cardinality

$$N^{-1} 2^{N-1} \beta e^{-\beta} (1 + o(1)), \text{ as } N \rightarrow \infty,$$

for a prefix of the form $10 \dots 0$.

Note that the above results indicate that codes avoiding one address sequence represent an exponentially large family of binary sequences. We prove a similar result for the case of 4-ary sequences that avoid a set of m mutually uncorrelated sequences. To establish the claim, we need the following definitions. Let $g(0), g(1), \dots$, be an integer sequence over a finite alphabet. Define the generating function of the sequence

$$G(z) = \sum_{N=0}^{\infty} g(N) z^{-N}.$$

Theorem 6. Suppose that $\{X_1, \dots, X_m\}$ is a set of mutually uncorrelated sequences of length n over the alphabet $\{\mathbf{A}, \mathbf{T}, \mathbf{C}, \mathbf{G}\}$. Let $f(N)$, with $f(0) = 1$, be the number of strings of length N over $\{\mathbf{A}, \mathbf{T}, \mathbf{C}, \mathbf{G}\}$ that do not contain substrings in $\{X_1, \dots, X_m\}$. Then

$$F(z) = \frac{z^N}{m + (z - 4)z^{N-1}},$$

where $F(z)$ is the generating function of the sequence $\{f(N)\}$.

Proof of Theorem 6. The result is a direct consequence of Theorem 4.1 of [17]. For $1 \leq i \leq m$, let $f_i(n)$ denote the number of strings of length n over $\{\mathbf{A}, \mathbf{T}, \mathbf{C}, \mathbf{G}\}$ that contain no element of $\{X_1, \dots, X_m\}$, except for a single copy of X_i at the right-hand side of the string. Let $F_i(z)$ be the generating function of $f_i(n)$. Then, we have the following system of equations that holds for the two sets of aforementioned functions:

$$\begin{aligned} (z - 4)F(z) + zF_1(z) + \dots + zF_m(z) &= z \\ F(z) - z(X_1 \circ X_1)_z F_1(z) - z(X_2 \circ X_1)_z F_2(z) - \dots - z(X_m \circ X_1)_z F_m(z) &= 0 \\ &\vdots \\ F(z) - z(X_1 \circ X_m)_z F_1(z) - z(X_2 \circ X_m)_z F_2(z) - \dots - z(X_m \circ X_m)_z F_m(z) &= 0 \end{aligned} \quad (2.2)$$

By using the fact that $(X_i \circ X_i)_z = z^{n-1}$, for $1 \leq i \leq m$, and $(X_i \circ X_j)_z = 0$, for $1 \leq i \neq j \leq m$, one can show that

$$F(z) = z^n F_1(z) = \dots = z^n F_m(z). \quad (2.3)$$

The result follows by replacing (2.2) into the first line of (2.3).

As the dominant pole of the generating function is close to 4, the number of sequences avoiding a set of mutually uncorrelated sequences grows roughly as 4^n .

Proof of Theorem 3. Since P is self-uncorrelated, we need to show that this string is not contained in the output of $\text{CodePSC}(P, \ell, x)$, where the output of $\text{CodePSC}(P, \ell, x)$ equals

$$\text{CodePSC}(P, \ell, x) = P^{t_1-1} \bar{p}_{t_1, s_1} \dots P^{t_r-1} \bar{p}_{t_r, s_r} \theta_{t_0}(\cdot),$$

for some input $\theta_{t_0}(\cdot)$, and $1 \leq t_0, t_1, \dots, t_r < n$. Consequently, if P is a substring of the output of $\text{CodePSC}(P, \ell, x)$, then the last symbol of P (recall that we assumed this symbol to be \mathbf{G}) has to appear in one of the following three positions:

- The symbol appears in P^{t_i-1} , for a unique $1 \leq i \leq r$: In this case, there exists a suffix of P appearing as a prefix of P^{t_i-1} . This contradicts our assumption that P is self uncorrelated.
- The symbol appears in \bar{p}_{t_i, s_i} , for a unique $1 \leq i \leq r$: This contradicts our assumption that $\bar{p}_{t_i, s_i} \neq \mathbf{G}$.
- The symbol appears in $\theta_{t_0}(\cdot)$: This contradicts our assumption that \mathbf{G} does not appear in $\theta_{t_0}(\cdot) \in \{\mathbf{A}, \mathbf{T}, \mathbf{C}\}^{t_0}$.

Therefore, the string P does not appear as a substring in the output of $\text{CodePSC}(P, m, x)$, which completes the proof.

Proof of Theorem 4. It suffices to show that the output of $\text{CodePSC}(P, \ell, x)$ is uniquely decodable. We use induction arguments to establish this result. For the basis step, by the definition of the output of CodePSC , it is straightforward to show that $\text{CodePSC}(P, \ell, x)$ returns the encoding $\theta_\ell(x)$, which represents a one-to-one mapping from $0 \leq x < 3^\ell$ to $\{\mathbf{A}, \mathbf{T}, \mathbf{C}\}^\ell$ whenever $\ell < n$. For the inductive step, we assume that the result is true for all $\ell < r$, as well as for all $r \geq n$, and show that it is consequently true for $\ell = r$.

For $\ell = r$, $\text{CodePSC}(P, \ell, x)$ returns

$$P^{t-1} \bar{p}_{t, s} \text{CodePSC}(P, \ell - t, b),$$

for some integer values s, b and for some $1 \leq t < n$, where $x = \left(\sum_{i=1}^{t-1} |\bar{P}_i| G_{n, \ell-i} \right) + (s-1) G_{n, \ell-t} + b$. Therefore x is uniquely decodable if and only if s, t and b are unique. Since sequences of the form $P^{t-1} \bar{p}_{t, s}$ are prefix-free one can uniquely identify both t and s . Moreover $\ell - t < r$, hence by the induction hypothesis it follows that b is also uniquely decodable from $\text{CodePSC}(P, \ell - t, b)$. Hence, x can be uniquely decoded.

Designation of primer	Sequence
B1-forward	5'AATTACTAAGCGACCTTCTC3'
B1-reverse	5'ACTTATTGCGACTTCTAAGG3'
gBlock-B1-reverse	5'CTTCATAACAACCTAACTGTGAC3'
B1-SU1-reverse	5'CGTGCACTCATAACCCATATTTCAAGAGCT AGCTATTCTCTCCCTTAAAAGTAAATGAC3'
B1-SD1-forward	5'GGGAGAGGAATAGCTAGCTCTTGAAATAT GGGTTATGAGTGCACGATCATCACATAAC3'
B2-forward	5'AACCTAACCATCTTCCTCTC3'
B2-reverse	5'AAACGATCCCCTGACAGAGC3'
gBlock-B2-forward	5'GAAGCACAGTGTTGCTGCGTG3'
B2-SU1-reverse	5'CAGCTTGATCCCATCTCAACCCTAATTC CATAACCGTCAGCGCAGTTGACTAGTCTC3'
B2-SD1-forward	5'CTGCGCTGACGGTTATGGAATTAGGGTT GAGATGGGATACAAGCTGATATGGGAAC3'
B3-forward	5'ATAATAGGCCTGATGATCTC3'
B3-reverse	5'AAGAAGAACCAGTAAGCAGC3'
B3-SU1-reverse	5'AACATCTACTCACTCTCAATCTAAGCTTGA ACTGTGTACACACCATCGCTCTTGACGCC3'
B3-SU2-forward	5'GTGTACACAGTTCAAGCTTAGATTGAGAGT GAGTAGATGTTGATGCGAGGCGAAAGATGT3'
B3-SD2-reverse	5'GACTTCCCCCTATAATCCATTAATGCTAG ATCAAGCCGCATATACTATGTTGCAAATAC3'
B3-SD2-forward	5'GCGGCTTGATCTAGCATTAAATGGATTA TAGGGGGGAAGTCGCTGCTGGTACTCTG3'

Table S2. List of primers for rewriting (editing) the blocks B1, B2 and B3. The primers for the gBlock method are listed separately for those used with the OE-PCR method. In the latter case, the labels of DNA fragments SU and SD stand for sample upstream and sample downstream. In OE-PCR, we linked two DNA fragments or three DNA fragments into the final PCR products; when two fragments were linked, the first fragment was labeled UP (U), while the second fragment was labeled DOWN (D); when three fragments were combined, the second fragment was labeled MIDDLE (M).

3 Address Sequences

Consider the following set of strings of length 20,

```
ACTAACTGTGCGACTGATGC
ACACTATCGAGCTGACACGT
AGTCAGCAGTAGTCAGTCAG
ACTGAGCTGAGCGTATATCG
ACTCAGCTACGACTCACATG
```

with GC content equal to 50%, i.e., 10 GC bases. The sequences are mutually uncorrelated and at Hamming distance exactly 10 from each other. The sequences do not exhibit secondary structures at room temperature, as verified by the mfold and Vienna packages. We used these addresses for a very small-scale, proof-of-concept random access/rewriting experiment of a 4 KB file.

In the large scale random access/rewriting experiment described in Section 5, we used different address sequences for the two flanking ends of the 1000 bps blocks. The sequences we synthesized include:

```
block 1: (CTCTCCAGCGAATCATTA, ACTTATTGCGACTTCTAAGG)
block 2: (CTCTCCTTCTACCAATCCAA, AAACGATCCCCTGACAGAGC)
block 3: (CTCTAGTAGTCCGGATAATA, AAGAAGAACCAGTAAGCAGC)
block 4: (CTCTTTCGCTGTGCACAAAA, AAATCGGAAATTCGTGTGCG)
block 5: (CTCTGCTGGAAATGTGTGAA, AATTCACGGTCCGAAACACC)
block 6: (CTCTGTTCTCCTTTCTCGT, TGTAGACGATTTGATTGGCG)
block 7: (CTCTAGCAACTTCCGCAAAT, ACGAGATTCATACCGGACCC)
block 8: (CTCTAGCTTCCCTATCCATA, TGCAGAAGAGGAGTGTACAGC)
block 9: (CTCTATAGGCTCTGGTATGT, TTAAACCCGCCCCGTACAGCC)
block 10: (CTCTCGCTCATCTCATGTTT, ACAGTACTTGCCCAATTGCG)
block 11: (CTCTGTACTCCGCTGAATCA, TAAACATTACAAGCCCCTCG)
block 12: (CTCTTCTTCCCTGACGATGT, AATACAACCTCTAACCACCC)
block 13: (CTCTTGATCCTACTGAGAAA, TTAATAGTTCCCGGCAGCCC)
block 14: (CTCTAGTGACGTGACAGGTA, TTAGAACGAACCAGTATAGC)
block 15: (CTCTACCTAAGGCCTTTGAA, TTGACCCATGAGCCAGCACC)
block 16: (CTCTACAGTAGTAAACTCGT, TGCTGAACTCTAATCTGTCC)
block 17: (CTCTGGGCGGCTGTACACAA, ATACACTCATAACACCTCGG)
block 18: (CTCTGCGATCACAAAAAGTT, ACAACTATACGTGTGCGACC)
block 19: (CTCTTTAGCACGAGTCCTAT, TGAACCCGTCGTGCTAATCG)
block 20: (CTCTAATACGCACGCCCAT, ATACGGGATACAATTAGGGC)
block 21: (CTCTGAGGCGTGGATATTTT, AATACATCCCTAAAAGCCGG)
block 22: (CTCTGCGTGTTTCATTCCATT, TGAGGATAGGATTAGTAAGG)
block 23: (CTCTAAGAATCTGACTGCAT, ATGTTAACTGAGTAAGGG)
block 24: (CTCTGATCGAACCCTGTCA, ACATGACCTACATAACGTCC)
block 25: (CTCTCTGGTGGCCTAAAAAT, AACAGAGATCAGAGCAGTGG)
block 26: (CTCTAGAGAAACGTTGAAGT, AACCCGTACTCACTATGCCG)
block 27: (CTCTGACGTCTACACAACAT, TTTGTAGATCCCAAGCATCG)
```

The pairs of sequences were used to flank the two ends of the data blocks. Only the addresses on the left were used for subsequent prefix-synchronized coding.

The sequences on the left-hand side of the pairing have “interleaved” $\{\mathbf{G}, \mathbf{C}\}$ and $\{\mathbf{A}, \mathbf{T}\}$ bases – for example, they all start with CTCT... This ensures a “GC balancing” property for the prefixes of the addresses.

4 Encoding and Decoding Example

In this section, we illustrate the encoding and decoding procedure for the short address string $P = \mathbf{AGCTG}$, which can easily be verified to be self-uncorrelated.

More precisely, we explain how to compute a sequence of integers $G_{n,1}, G_{n,2}, \dots, G_{n,7}$, described in the main body of the paper. As before, n denotes the length of the address string, which in this case equals five.

One has

$$(G_{n,1}, G_{n,2}, \dots, G_{n,7}) = (3, 9, 27, 81, 267, 849, 2715).$$

The algorithm $\text{CodePSC}(P, 8, 550)$ produces:

$$\begin{aligned} 550 &= 0 \times G_{5,7} + 550 \\ &\Rightarrow \text{CodePSC}(P, 8, 550) = \underline{\text{C}}\text{CodePSC}(P, 7, 550) \\ 550 &= 0 \times G_{5,6} + 550 \\ &\Rightarrow \text{CodePSC}(P, 7, 550) = \underline{\text{C}}\text{CodePSC}(P, 6, 550) \\ 550 &= 2 \times G_{5,5} + 0 \times G_{5,4} + 16 \\ &\Rightarrow \text{CodePSC}(P, 6, 550) = \underline{\text{AA}}\text{CodePSC}(P, 4, 16), \\ 16 &= 0 \times 3^3 + 1 \times 3^2 + 2 \times 3^1 + 1 \times 3^0 \\ &\Rightarrow \text{CodePSC}(P, 4, 16) = \underline{\text{ATCT}}, \\ &\Rightarrow \text{CodePSC}(P, 8, 550) = \underline{\text{CCAAATCT}} \end{aligned}$$

When running $\text{DecodePSC}(P, X)$ on the encoded output $X = \underline{\text{CCAAATCT}}$, the following steps are executed:

$$\begin{aligned} &\Rightarrow \text{DecodePSC}(P, \underline{\text{CCAAATCT}}) = 0 \times G_{5,7} \\ &+ \text{DecodePSC}(P, \text{CAAATCT}) \\ &\Rightarrow \text{DecodePSC}(P, \underline{\text{CAAATCT}}) = 0 \times G_{5,6} \\ &+ \text{DecodePSC}(P, \text{AAATCT}), \\ &\Rightarrow \text{DecodePSC}(P, \underline{\text{AAATCT}}) = 2 \times G_{5,5} + 0 \times G_{5,4} \\ &+ \text{DecodePSC}(P, \text{ATCT}) \\ &\Rightarrow \text{DecodePSC}(P, \underline{\text{ATCT}}) = 16 \\ &\Rightarrow \text{DecodePSC}(P, \text{CCAAATCT}) = 2 \times G_{5,5} + 16 = 550 \end{aligned}$$

5 Experimental Synthesis, Access and Rewrite of DNA Sequences

A total of 27 sequences of length 1000 bps each were designed to encode information retrieved from the Berkeley, Harvard, MIT, Princeton, Stanford, and UIUC Wikipedia page in 2014. Except for sequence #4, which was rejected due to the complexity of its secondary structure, all sequences were synthesized by IDT (Integrated DNA Technologies). In addition, 27 corresponding address primers were synthesized by the same company. The address sequences of the blocks are listed in Section 3.

As a proof of concept, we performed a number of selection and editing experiments. These include selecting individual blocks and rewriting one of its sections, selecting three blocks and rewriting three sections in each, two close to the flanking ends, and one in the middle. The edits involved information about the budget of the institutions at a given year of operation. Detailed information about the original sequences and their rewritten forms is given in the following sections.

Sequence identifier	Number of sequence samples	Length of the edited region (in bps)	Selection accuracy / readout error percentage	Description of editing method
B1-M-gBlock	5	20	5/5/0%	gBlock method
B1-M-PCR	5	20	5/5/0%	OE-PCR method
B2-M-gBlock	5	28	5/5/0%	gBlock method
B2-M-PCR	5	28	5/5/0%	OE-PCR method
B3-M-gBlock	5	41 + 29	5/5/0%	gBlock method
B3-M-PCR	5	41 + 29	5/5/0%	OE-PCR method

Table S3. Selection, rewriting and sequencing results. Each rewritten 1000 bps sequence was ligated to a linearized pCRTM-Blunt vector using the Zero Blunt PCR Cloning Kit and was transformed into *E. coli*. The *E. coli* strains with correct plasmids were sequenced at ACGT, Inc. Sequencing was performed using two universal primers: M13F_20 (in the reverse direction) and M13R (in the forward direction) to ensure that the entire blocks of 1000 bps are covered.

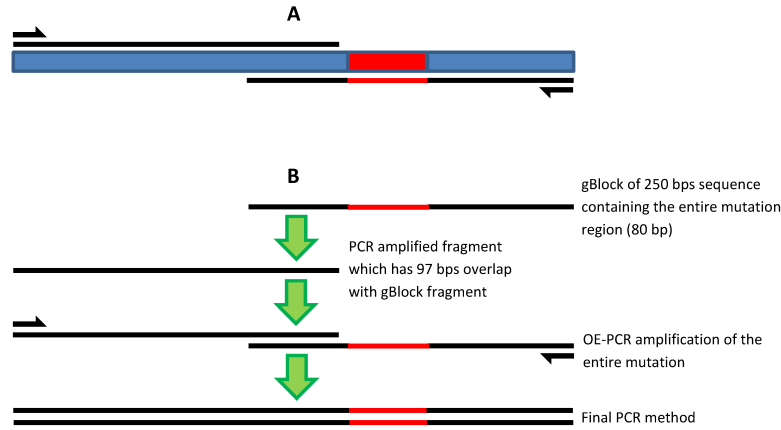


Fig. S1. A) Schematic depiction of the editing method using gBlocks. B) Detailed description of the generation of the mutation. Four sequences (ranging in length from 177 to 588 bps) containing the entire edit region were gBlock synthesized from IDT. The remaining parts of the 1000 bps sequences were PCR amplified. A homology in at least 30 bps between the flanking end sequence of the blocks and the corresponding end of the gBlock fragment was created. By one OE-PCR, the desired edits were generated in a one-pot matter.

We denoted the blocks on which we performed selection and editing by B1, B2, and B3. The primers used for performing the edits in the blocks are listed in Table S2. Note that two primers were synthesized for each rewrite, for the forward and reverse direction. In addition, two different editing (mutation) techniques were used, gBlock and Overlap-Extension (OE) PCR; gBlocks are double-stranded genomic fragments that are frequently used as primers, for gene construction or for mediated genome editing. An illustration of editing via gBlocks is shown in Fig. S1. On the other hand, OE-PCR is a variant of PCR used for specific DNA sequence editing via point mutations or splicing. An illustration of the procedure is given in Fig. S1. To demonstrate the plausibility of a cost efficient method for editing, OE-PCR was used with general primers (≤ 60 bps) only. For edits shorter than 40 bps, the mutation sequences were designed as overhangs in primers. Then, the three PCR products were used as templates for the final PCR reaction involving the entire 1000 bps rewrite.

All 27 linear 1000 bps fragments were mixed, and the mixture was used as a template for PCR amplification and selection of the B1, B2 and B3 sequences. The results of selection are shown in Fig S2, where three banks of size 1000 bps are depicted. These banks indicate that sequences of the correct length were isolated. Subsequent sequencing confirmed that the sequences were indeed the user requested B1, B2 and B3 strands. A summary of the experiments performed is provided in Table S3.

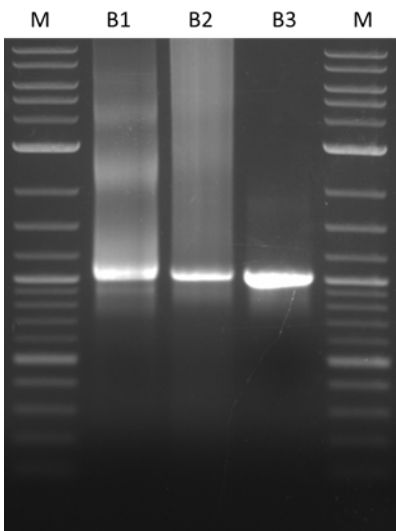


Fig. S2. PCR of 1000 bps sequences-B1, B2, B3 from a mixture of 26 sequences.

5.1 B1 mutation B1-M synthesis

The unedited B1_original (B1) sequence is of the form:

```

AATTACTAAGCGACCTTCTCGGATAGAACGCTTAGTTGGTGCGTTGACAT
GCTCGAACTGATCATCGGTCACTTGCATTCAATTATTGATTGTTGAGTTGA
GAAGCGCATTGGTGCTACTCGTTGCTGGGTCATTTTCGGCGAGAGAAACA
GTTCACTGTGGCGTGATGTTTTGAAATGAGGGAGAGTTCTCTTAACTGCA
GTTGGAGTTCAGTATACTCGGGATAGTGTAACAGAGGGAGGCGGATGTGT
GTATTGATGTGAAGTCTTTCACGTGCGGGCTAGGTCGTAATGACGGGTCG
GGAAGTATTCATTGGCGCAATAGTGATTTTGATGAATGATGGATAGAACG
CTTAAAGGGAACTATATAGTTCAAAGCTCGTCGGCGGTGTCGAGGATGT
ATAGGGGTTAATGAATGGTGGAAGTACTTATACTATAGATTGGACTGGT
GGTATGAGAACTTCACTAATTATTGACGTCACAGTTAGTTGTTATGAAGT
GATAATATGAATCGAGCGCAACAGGACTAGTCATTTACTTTTAAGGGAGA
GGAATAGCTAATCTCAAATTTTTTTTATGTGAGTGCACGATCATCACATA
ACATAGGAGGCGATGAGACAGCGACTCAATCTGACTAATTCATTATAGGA
GTTATATGAAGAGTTCGGAACGAAGCTAGCGCTTTCGCACAATGCGAGGG
ATAAGAGCGGGTGCAGAGCGAAGGGTGTGAAATTGATGGTGGATAAGAAC
TTCGCACAGTACTAGCTAGTGGGGAGAGACTTCTATGAATTCGGAGGGAT
ACTTGATATTGATATGGGGGGATGGCGCTATTAAGCGCAGAGCGTAAGTG
CGCTTCAAATCGAACATTGTGTAGCTAAGCAATAGAGAAATGTGGGGATT
GAGCAGTTCGTATCGGTTTCGCATGACATACTTGGGAAAAATGGCAGCTTGT
TTAAGCTAACTGGATGAAAGGGAGGAAAAAATTATTGCGACTTCTAAGG

```

where the bases written in red represent the regions we edited.

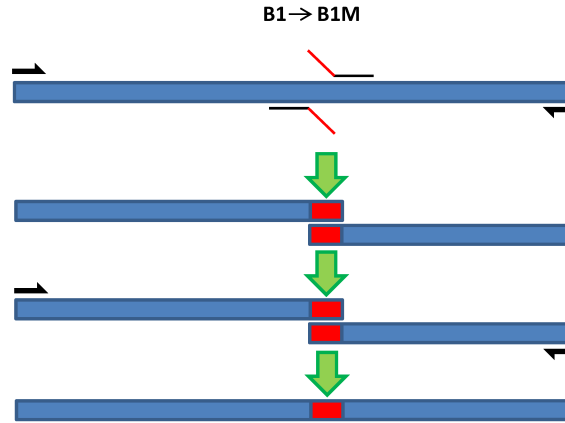


Fig. S3. Illustration of the process of generating the B1 edit/mutation using general primers.

The edited B1_mutation (B1_M) sequence reads as:

```

AATTACTAAGCGACCTTCTCGGATAGAACGCTTAGTTGGTGCGTTGACAT
GCTCGAACTGATCATCGGTCACTTGCATTCAATTATTGATTGTTGAGTTGA
GAAGCGCATTGGTGTCACTCGTTGCTGGGTCATTTTCGGCGAGAGAAACA
GTTCACTGTGGCGTGATGTTTTGAAATGAGGGAGAGTTCTCTTAAGTGA
GTTGGAGTTCAGTATACTCGGATAGTGTAACAGAGGGAGGCGGATGTGT
GTATTGATGTGAAGTCTTTCACGTGCGGGCTAGGTCGTAATGACGGGTCG
GGAAGTATTCATTGGCGCAATAGTGATTTTGATGAATGATGGATAGAACG
CTTAAAGGGAACTATATAGTTCAAAGCTCGTCGGCGGTGTCGAGGATGT
ATAGGGGTTAATGAATGGTGAAGTACTTATACTATAGATTGGACTGGT
GGTATGAGAACTTCACTAATTATTGACGTCACAGTTAGTTGTTATGAAGT
GATAATATGAATCGAGCGCAACAGGACTAGTCATTTACTTTTAAGGGAGA
GGAATAGCTAGCTCTTGAATATGGGTTATGAGTGCACGATCATCACATA
ACATAGGAGGCGATGAGACAGCGACTCAATCTGACTAATTCATTATAGGA
GTTATATGAAGAGTTCGGAACGAAGCTAGCGCTTTCGCACAATGCGAGGG
ATAAGAGCGGGTGCAGAGCGAAGGGTGTGAAATTGATGGTGGATAAGAAC
TTCGCACAGTACTAGCTAGTGGGGAGAGACTTCTATGAATTCGGAGGGAT
ACTTGATATTGATATGGGGGATGGCGCTATTAAGCGCAGAGCGTAAGTG
CGCTTCAAATCGAACATTGTGTAGCTAAGCAATAGAGAAATGTGGGGATT
GAGCAGTTCGTATCGGTTTCGCATGACATACTTGGGAAAATGGCAGCTTGT
TTAAGCTAACTGGATGAAAGGGAGGAAAAAATTATTGCGACTTCTAAGG

```

with rewrites listed in red.

5.1.1 The gBlock method

Since a gBlock of length longer than 500 bps was needed, it was more costly to synthesize the gBlock and perform rewriting than to directly re-synthesizing the whole block. Hence, the gBlock method was not used in this case.

5.1.2 The OE-PCR based method

One pair of primers was designed to PCR amplify the first portion of the sequence B1-M. For the forward direction, the primer was

5'AATTACTAAGCGACCTTCTC3'

while for the reverse direction, the primer was

5'CGTGCACTCATAACCCATATTTCAAGAGCTAGCTATTCCTCTCCCTTAAAAGTAAATGAC3'.

The second part of the sequence was PCR amplified by using the forward direction primer

5'GGGAGAGGAATAGCTAGCTCTTGAAATATGGGTATGAGTGCACGATCATCACATAAC3'

and reverse direction primer

5'ACTTATTGCGACTTCTAAGG3'.

Both PCR reactions used the sequence B1 as template. Two such PCR products are shown in Fig. S4, indicating that the correct length products were isolated in each reaction.

OE-PCR was performed in a 50 ul reaction volume containing the two aforementioned PCR products without primers for the first 5 cycles and the products with primers (B1 primers in Table S2) for the later 30 cycles. A single bank with correct size of 1000 bps was obtained (see Fig. S4).

5.2 B2 mutation B2-M synthesis

The unedited B2_original (B2) sequence is of the form:

```
AACCTAACCATCTTCCTCTCGATTGGAGCAGATTGGTATTATTCTAGTC
GTCGAGACTAGTCAACTGCGCTAGTTTGTGTTCAAAAATAAGAGTATGA
GATACAAGCTGATATGGGAACCTAATTACGAAGCACAGTGTTGCTGCGTG
GACTTGTGAAGTAGGGTGTGAGATAAGAATGATAGCGAACGCAGCGTATG
GCTGAAGTGCTGGGCATATTGTGGTGTGGACATCTCAAAGTCTATGAAGA
TTGGTAATAGGATGGTCTCTCGGGTCTCAAACCTTCGTCAGGCAGCATTGT
GCATGCGAGTGATTGAAAGGGAGGGTAAGGGTTATTAATAGAAAAGACTT
ACAGGCGTTGGTATGATTCAAGATCGCAAGAATCGTGTGAGCTTGAGGAC
TAAATAGTTTTAAAGAAATAGGAATAGTTGTAATTTAAGGAGCGTGGCACG
GATGGATCAGCGTGTCAACGGAACGCGCATTTGGGAGTTTTATGTAAAGT
GAGCAGACTAAGGTGAAATTCAATAGTCTCTATCGTTCGAGGGTTATTGC
TAGGGGAGACTTTGAGTGAGTGGTAATTTGAAGCAGTATACGTAACTTT
TTCGATTCTTAGTGGCAGTTACTCTGAATTTTAGTGTGAGCAGAGTGTGA
TAAATAGAGAGATACGAGGTCGACACGGCTGTTGGGGGCACTTAACAGTA
GGGGGTTGATGCTGGCGGACACTAAAGGATTTTTGAAGGGGATTGTTGGC
GACTCACATCTAAGTGGTATTGCGGGCTCTATGAGAATCTGCTCGAGTCA
TCTAGGTTGAGGAAGAGGGGAGATTCTCGTTAAAGACAGTACATATTTTC
GCATACTTCTTAACGTGGAGTATGAATGTCAATGGTGGGAGATATGGGTG
GAGGGATTTCACTTCACTGCATATGTACGCTCAGGAGCGCGAACGAATCAT
AAAACCTATTGTAATATATTGATAGATAAAGAAACGATCCCCTGACAGAGC
```

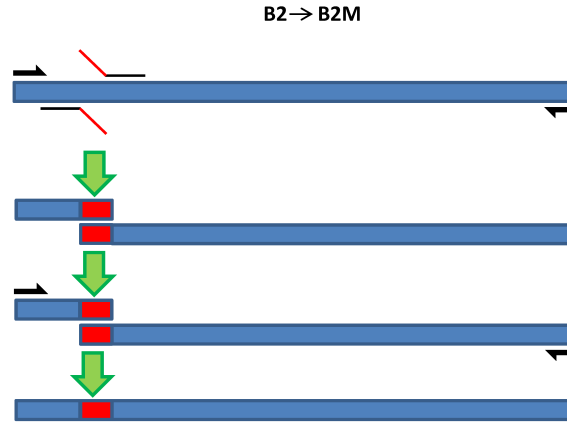


Fig. S4. A schematic depiction of the process of generating the B2 mutation using standard 60 bps primers.

The edited B2_mutation (B2_M) sequence is:

```

AACCTAACCATCTTCCTCTCGATTTGGAGCAGATTGGTATTATTCTAGTC
GTCGAGACTAGTCAACTGCGCTGACGGTTATGGAATTAGGGTTGAGATGG
GATACAAGCTGATATGGGAACCTAATTACGAAGCACAGTGTGCTGCGTG
GACTTGTGAAGTAGGGTGTGAGATAAGAATGATAGCGAACGCAGCGTATG
GCTGAAGTGCTGGGCATATTGTGGTGTGGACATCTCAAAGTCTATGAAGA
TTGGTAATAGGATGGTCTCTCGGGTCTCAAACCTTCGTCAGGCAGCATTGT
GCATGCGAGTGATTGAAAGGGAGGGTAAGGGTTATTAATAGAAAAGACTT
ACAGGCGTTGGTATGATTCAAGATCGCAAGAATCGTGTGAGCTTGAGGAC
TAAATAGTTTAAAGAAATAGGAATAGTTGTAATTTAAGGAGCGTGGCACG
GATGGATCAGCGTGTCAACGGAACGCGCATTGGGAGTTTTATGTTAAGT
GAGCAGACTAAGGTGAAATTCAATAGTCTCTATCGTTCGAGGGTTATTGC
TAGGGGAGACTTTGAGTGAGTGGTAATTTTGAAGCAGTATACGTAACTTT
TTCGATTCTTAGTGGCAGTTACTCTGAATTTTAGTGTGAGCAGAGTGTGA
TAAATAGAGAGATACGAGGTCGACACGGCTGTTGGGGCACTTAACAGTA
GGGGTTGATGCTGGCGGACACTAAAGGATTTTGAAGGGGATTGTTGGC
GACTCACATCTAAGTGGTATTGCGGGCTCTATGAGAATCTGCTCGAGTCA
TCTAGTTGAGGAAGAGGGGAGATTCTCGTTAAAGACAGTACATATTTTC
GCATACTTCTTAACGTGGAGTATGAATGTCAATGGTGGGAGATATGGGTG
GAGGGATTTCAATCACTGCATATGTACGCTCAGGAGCGCGAACGAATCAT
AAAACCTATTGTAATATATTGATAGATAAAGAAACGATCCCCTGACAGAGC

```

where, as before, red letters were used to indicate the rewritten region.

5.2.1 The gBlock method

A 177 bps sequence, containing the entire edited region and the B2 string, was gBlock synthesized by IDT. Another part of B2 was PCR amplified using the forward primer

5'GAAGCACAGTGTGCTGCGTG3'

and reverse primer

5'AAACGATCCCCTGACAGAGC3'

The B2 sequence served as a template. See Fig. S4 for an illustration.

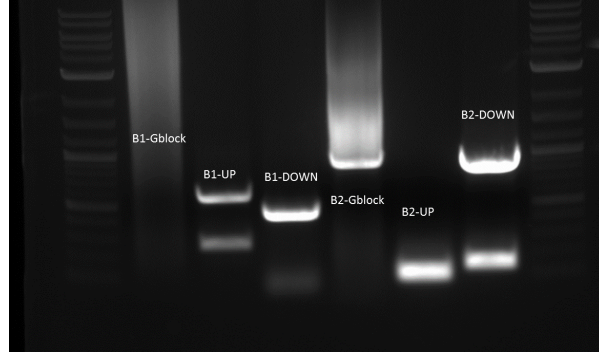


Fig. S5. PCR products of B1 and B2.

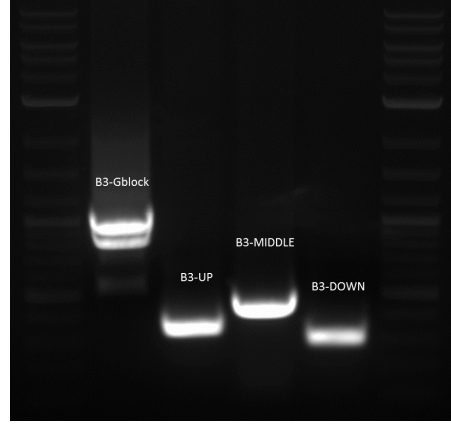


Fig. S6. PCR products of B3.

5.2.2 The OE-PCR based method

Over extension PCR (OE-PCR) was performed in a 50 ul reaction volume containing the above 177 bps gBlock product and PCR products without primers for the first 5 cycles and with B2 forward and reverse primers listed in Table S2 for the subsequent 30 cycles.

The PCR product was deposited on a gel substrate and the correct 1000 bps band was obtained as shown in Fig. S5. One pair of primers was designed to PCR amplify the first part of the sequence B2-M, with forward primer

5'AACCTAACCATCTTCCTCTC3'

and reverse primer

5'CAGCTTGTATCCCATCTCAACCCTAATTCCATAACCGTCAGCGCAGTTGACTAGTCTC3'.

The second part was PCR amplified by the forward primer

5'CTGCGCTGACGGTTATGGAATTAGGGTTGAGATGGGATACAAGCTGATATGGGAAC3'

and reverse primer

5'AAACGATCCCCTGACAGAGC3'.

Both PCRs used B2 as a template. Two PCR products are shown in Fig. S5.

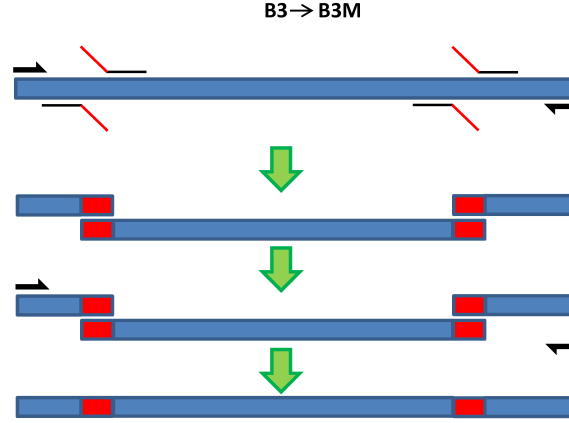


Fig. S7. Scheme for generating the B3 edits using standard 60 bps primers.

5.3 B3 mutation B3-M synthesis

The unedited original B3 sequence equals:

```

ATAATAGGCCTGATGATCTCGATGGATGCGCGTCACTCGAGTGCGGTAGG
CACGTCTCAGGTGATAAGTGATTGTGATTGTAGGTGAAGGGGGTAGAAAT
GATTGAGGAACTTGTGTACTCGTTACACGTGATAGGGTTTGATCGGCGG
TGGA AAAAATTAGGGATGGGGATAAGATTATGGGATCGTTCTCAATAATTG
TTACGATATCGTTGTTACACAGTTGTTACGCTACGACGTCATCGATAAAG
GTGGGTATGTGGGGTACTATACTCTTGGGGGCGTACAAGAGCGATGGTT
GGTCGGATTGAAATTAAAAGCATTAAAGAGGTTAATTTATAGATGCGAGGC
GAAAGATGTGAGCGCAAGTAAAGGAAACGCGAGCAAGTGATTGTTACTAA
TTATATTAGGAGGTGATGAGGAGCGTGGTTATCTTATTGGGCGAGCTGCA
GCGAATTCTAGATTTCTTCGAGTTACAGTCGTAGTGATGTATATAGAGTG
GATGCGCACATTATTACATATATCGTCGAATTGGATTAGACGCAAAGAAA
ATGCGGCATTGTAATGGGTTGTGTA AAATTGAGCGTGGTTATCTTGTCAT
GACATAGTAAAAGTTGCTCAATTGATTGAAGCTCGATTAGGAGAAGTAAT
TTGAAAAAAGGATAGACTAGGACTCAACGAGGAACGGGTATTTGCAACAT
AGTATATGCGGTCTTAATCGGAGGGTAATGTTATTTGTGTGGAAGTCGCT
GCTGGTACTCTGGGCGTTTAGGATGAATCTTCGAAACTAGGCTTTGTCAG
AGATAGTTTGTGTTGTAAGAAGAATCAGGAAACGGTAACAGAGAATAAATG
AATTAACGTAGCAAGATTTTCGTCTTTCTGGAGATGAGAAGGTGTAGTTGA
GGAGTCGACGTTCTTTACGGAGGTGGGAGATTGGTTTTTGGCAGTACTTCG
TTAAATACACTAAAAAATTTGATAATGTAGAAGAAGAACCAGTAAGCAGC

```

The edited sequence B3_M mutation sequence is:

```
ATAATAGGCCTGATGATCTCGATGGATGCGCGTCACTCGAGTGCGGTAGG
CACGTCTCAGGTGATAAGTGATTGTGATTGTAGGTGAAGGGGGTAGAAAT
GATTGAGGAACTTGTGTACTCGTTACACGTGATAGGTTTGATCGGCGG
TGAAAAATTAGGGATGGGGATAAGATTATGGGATCGTTCTCAATAATTG
TTACGATATCGTTGTTACACAGTTGTTACGCTACGACGTCATCGATAAAG
GTGGGTATGTGGGGTACTATACTCTTGGGGCGGTACAAGAGCGATGGTG
TGTACACAGTTCAAGCTTAGATTGAGAGTGAGTAGATGTTGATGCGAGGC
GAAAGATGTGAGCGCAAGTAAAGGAAACGCGAGCAAGTGATTGTTACTAA
TTATATTAGGAGGTGATGAGGAGCGTGTTATCTTATTGGGCGAGCTGCA
GCGAATTCTAGATTTCTTCGAGTTACAGTCGTAGTGATGTATATAGAGTG
GATGCGCACATTATTACATATATCGTCGAATTGGATTAGACGCAAAGAAA
ATGCGGCATTGTAATGGGTTGTGTAATAATTGAGCGTGTTATCTTGTCAT
GACATAGTAAAAGTTGCTCAATTGATTGAAGCTCGATTAGGAGAAGTAAT
TTGAAAAAAGGATAGACTAGGACTCAACGAGGAACGGGTATTTGCAACAT
AGTATATGCGGCTTGATCTAGCATTAAATGGATTATAGGGGGGAAGTCGCT
GCTGGTACTCTGGGCGTTTAGGATGAATCTTCGAACTAGGCTTTGTCAG
AGATAGTTTGTGTTGTAAGAAGAATCAGGAAACGGTAACAGAGAATAAATG
AATTAACGTAGCAAGATTTCTGCTTTTCTGGAGATGAGAAGGTGTAGTTGA
GGAGTCGACGTTCTTTACGGAGGTGGGAGATTGGTTTTGGCAGTACTTCG
TTAAATACACTAAAAAATTTGATAATGTAGAAGAAGAACAGTAAGCAGC
```

5.3.1 The Gblock method

Two sequences, the 560 bps sequence containing the first mutation region and the second 560 bps sequence containing the second mutation region, were gBlock synthesized by IDT. There was a 60 bps overlap between the two gBlocks.

5.3.2 The OE-PCR method

OE-PCR was performed in a 50 ul reaction volume containing the above two 560 bps gBlock products without primers for the first 5 cycles and additional B3 forward and reverse primers listed in Table S2 for the subsequent 30 cycles. The PCR product was deposited on a gel substrate and the correct 1000 bps band was obtained.

One pair of primers was designed to PCR amplify the first part of the sequence B2-M, using

5'ATAATAGGCCTGATGATCTC3'

in the forward direction and

5'AACATCTACTACTCTCAATCTAAGCTTGAAGTGTGTACACACCATCGCTCTTGACGCC3'

in the reverse direction.

The second part was PCR amplified in the forward direction by using the primer

5'GTGTACACAGTTCAAGCTTAGATTGAGAGTGAGTAGATGTTGATGCGAGGCGAAAGATGT3'

and in the reverse direction by using the primer

5'GACTTCCCCCTATAATCCATTAATGCTAGATCAAGCCGCATATACTATGTTGCAAATAC3'.

The third part was PCR amplified by the forward direction primer

5'GCGGCTTGATCTAGCATTAAATGGATTATAGGGGGGAAGTCGCTGCTGGTACTCTG3'

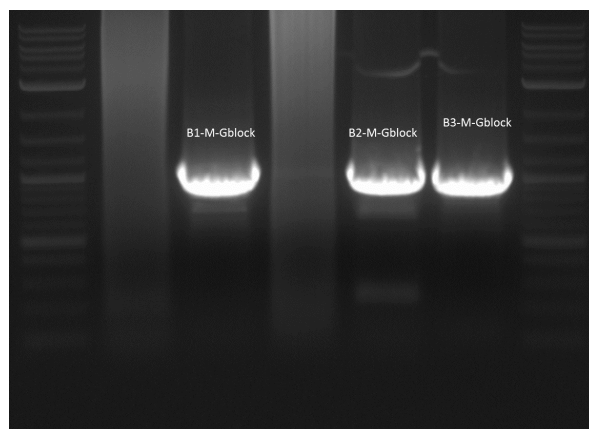


Fig. S8. The generated PCR products of 1000 bps edits from the gBlock method, involving B1-gBlock, B2-gBlock and B3-gBlock.

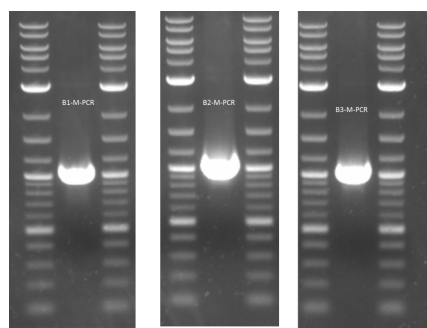


Fig. S9. The generated PCR products of 1000bps sequence editing for the OE-PCR based method, and sequences B1-PCR, B2-PCR and B3-PCR.

and reverse direction primer

5'AAGAAGAACCAGTAAGCAGC3'.

All three PCRs used the sequence B3 as the template. All three PCR products are shown in Fig. S8.

OE-PCR was performed in a 50 ul reaction volume containing the above three PCR products without primers for the first 5 cycles and with B3 primers listed in Table S2 for the subsequent 30 cycles. A single bank of correct size 1000 bps was obtained (See Fig. S9).

Correctness of the synthesized edited regions was confirmed via DNA Sanger sequencing as follows. The PCR products of the gBlock method and the OE-PCR method were named B1-M-gBlock, B2-M-gBlock, B3-M-gBlock and B1-M-PCR, B2-M-PCR, B3-M-PCR, respectively. All final mutations/edits of PCR products were purified using the QiaGen Gel Purification Kit. The purified 1000 bps edited sequences were blunt-ligated to the vector named pCRTM-Blunt (Fig. S10) using the Zero Blunt PCR Cloning Kit and following the manufacturers' protocol. Five colonies of each PCR-Blunt-mutation were sent to ACTG, Int. Sequencing was performed using two universal primers: M13F_20 (for the reverse direction) and M13R (for the forward direction). Bi-directional sequencing was performed in order to ensure that the entire 1000 bps block was completely covered.

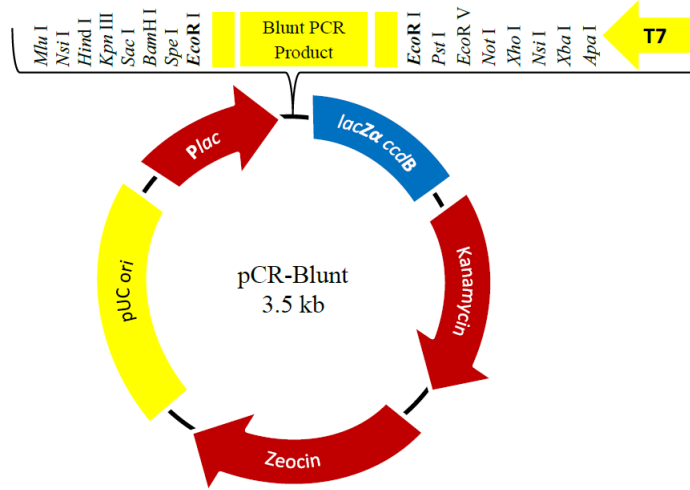


Fig. S10. Map and features of PCR-Blunt vector (Life technologies).

6 Hybrid DNA-Based and Classical Storage

In our small-scale experiments, Sanger sequencing produced two erroneous symbols in one strand which we were able to correct using prefix matching. One possible problem that may arise in large scale DNA-storage systems involving millions of blocks is erroneous sequencing which may not be corrected via prefix matching. In current High Throughput Sequencing technologies, such as Illumina HiSeq or MiSeq, the dominant sources of errors are substitutions. Due to our word grouping scheme, such substitution errors cannot cause catastrophic error propagation, but may nevertheless accumulate as the number of rewrite cycles increases. In this case, prefix matching may not suffice to correct the errors and more sophisticated coding schemes need to be used. Unfortunately, adding additional parity-check symbols into the prefix-encoded data stream may cause problems as the parities may violate the prefix properties and dis-balance the GC content. Furthermore, every time rewriting is performed, the parity-checks will need to be updated, which incurs additional cost for maintaining the system. A simple solution to this problem is a hybrid scheme, in which the bulk of the information is stored in DNA media, while only parity-checks are stored on a classical device, such as flash memory. Given that the current error-rate of short-read sequencing technologies roughly equals 1%, the most suitable codes for performing this type of coding are low-density parity-check codes [19]. These codes offer excellent performance in the presence of a large number of errors and are decodable in linear time.